STUDY OF MARKERS FOR REGULATORY ELEMENTS IN HUMAN GENOME

by

Vatsal Agarwal

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland
October 2013

Abstract

Most genetic traits and diseases in humans from height to cancer or sudden cardiac death do not follow Mendelian principles but originate from complex combinatorial effects of multiple genes with possibly multiple variants. Most of these variants lie within non-coding regions of the genome such as promoters, enhances or insulators, which regulate the expression levels of genes. Numerous algorithms predict the likely location of these regulatory regions using biological features such as conservation, transcription factor binding, deoxyribonuclease I (DNaseI) hypersensitivity, and others. The first part of the thesis presents a software to compile such annotations and visualize them in a customizable manner. The second part discusses the distribution of one of these features, DNaseI sensitivity, across the human genome.

In the first part, we developed a software and used it to study the *NOS1AP* (NO-synthase adapter protein) gene locus and the beta-globin gene locus. Since, single nucleotide polymorphisms (SNPs) at *NOS1AP* locus are known to affect the electro-cardiographic QT-interval, we collected the corresponding data from a genome-wide association study. We plotted the genetic effect and frequency of these SNPs across the length of the *NOS1AP* locus, along with genes and other functional annotations from various public databases including RefSeq, University

of California Santa Cruz (UCSC) Genome Browser, TRANSFAC, and the Encyclopedia of DNA Elements (ENCODE) project. We also added SNPs from the 1000 Genomes project to increase the available number of variants to analyze. We observed a lack of known annotations at almost all variants, which led to the following possibility: although particular regions of the human genome may not be significant enough to be designated as regulatory regions, there may still be weak sites affecting overall gene expression. This was the motivation to study the distribution of DNaseI sensitivity across the human genome, which forms the second part of the thesis.

In the second part, we modeled DNaseI sensitivity, a marker for chromatin accessibility and regulatory elements, using data collected by the University of Washington (UW) as part of the ENCODE project. We used Gamma-weighted Poisson distribution as our model and normal Poisson distribution as noise. Maximum-likelihood estimation fitting over the entire genome as well as over individual chromosomes, across different cell lines, indicated that most of the human genome is inactive, and the remainder has generally very low DNaseI sensitivity. Only a very small fraction of the genome (<1%) is DNaseI hypersensitive.

Primary reader: Dr. Aravinda Chakravarti (Advisor).
Secondary readers: Dr. Michael Beer, Dr. Liliana Florea.

Acknowledgement

I am grateful to Dr. Aravinda Chakravarti, my mentor and advisor for not only providing me the opportunity and guidance to do this project but also teaching me ways and ethics to conduct proper research.

I would like to thank Dr. Ashish Kapoor for his thorough inputs in this project and thesis as well as his personal help and support for past year and a half.

I would also like to take this opportunity to thank all the members of my lab and my friends for suggesting ideas for some aspects of the project from time to time and keep me motivated to complete the thesis.

Finally, I am truly grateful to my parents for encouragement to pursue graduate studies and their endless love and support in all aspects of my life, without which it would not be possible for me to complete this thesis.

Table of Contents

List of figures

Chapter 2

Chapter 3

Chapter 1

Introduction

1.1 Non-mendelian genetics and complex traits

1.1.1 Overview and application

Physical traits studied in early genetics were simple and monogenic in nature, following Mendelian principles where a significant mutation in one of the genes caused a distinguished phenotype or a disease. Fischer's model extended this logic to multiple genes and quantitative trait loci where expression of multiple genes would have additive effect on the phenotype [1]. However, these principles account for a small number of traits.

Improvement in sequencing technologies for sequencing of exomes to complete genomes, coupled with steeply falling prices for sequencing, has provided the scientific community with a huge amount of genetic data to analyze the correlation of sequence variation to human genetic traits and diseases. Genome-wide association studies (GWAS) have been performed

for many traits and diseases, but these are able to explain only a small portion of observed phenotypic variation [2]. Moreover, GWAS are based on the principle of Linkage Disequilibrium [3, 4], and hence, only highlight the target loci rather than identifying the causal variation.

However, data from GWAS of over 240 traits and diseases, identifying over 3500 associated SNPs, shows that about 88% of these SNPs lie within non-coding region of the genome [5]. These non-coding variants are hypothesized to lie in regulatory regions of the genome, which regulate gene expression. So, the aim to identify the causal variation would be a step closer if we could locate the regulatory regions in the genome. Unfortunately, there are many classes of regulatory elements that have significantly different structure and function. Promoters are responsible for initiating and regulating transcription processes and lie upstream of the gene on the same strand; enhancers increase the pace of transcription whereas suppressors decrease the speed, but both of these may lie far from the gene they regulate; insulators act as an impermeable wall to prevent the effect of certain enhancers and suppressors beyond a certain region; transcription factor binding sites, as the name suggests, are locations that are bound by transcription factors.

Although there is no universal method or marker to identify all regulatory elements, we know of few biological properties and functional annotations that hint toward the locations of regulators. Conservation is considered one of these. If a region of the genome is conserved across species, it may have an important role to play. Binding sites for transcription factors also provide an important resource in this direction [6]. Openings of chromatin found by DNase I hypersensitive sites (DHS) are generic markers for several classes of regulatory elements [7, 8].

1.2 Dissertation outline

1.2.1 Software to compile and visualize various known functional annotations

Numerous mathematical algorithms model one of the several functional annotations to estimate regulatory regions. For instance, JASPER [9] and TRANSFAC [10] use transcription-factor binding, whereas as part of the Encyclopedia of DNA elements (ENCODE) Project[11], University of Washington [12] and Duke University [13] employ DHS in their algorithms. In this chapter, we discuss a software we developed to analyze regions with multiple publicly available annotations, by visualizing them along the length of a chromosome.

This software enabled us to gain insights by looking at the plots and would be useful to researchers to study specific regions of genome in detail.

1.2.2 Genome-wide modeling of DNase I sensitivity

Inconsistencies in annotation from different sources and lack of marked regulatory regions at expected locations led us to hypothesize the presence of weaker sites which could not pass algorithmic thresholds. In this chapter, we studied the distribution of regulatory regions by modeling DNase I sensitivity as a Gamma distribution across the human genome, in various cell lines.

This model gives consistent results among replicates and shows expected behavior in chromosomal variation. It successfully helps us understand the distribution of DNase sensitivity. We inferred that roughly 90% of the genome is inactive, 9.9% has low sensitivity and forms weaker sites and only about 0.1% of the genome is hypersensitive in nature.

Chapter 2

Annotation visualization software

2.1 Introduction

Several mutations in non-coding portions of the genome are responsible for many known complex traits and are capable of causing diseases [5]. These mutations lie in regulatory regions and affect gene expression levels. Hence, it is important to identify parts of genomes which act as regulators. Different regulatory elements may be surveyed in different applications, some of which may be involved in specific cell types. Hence, there is yet no universal method for their identification. However, several types of features including transcription-factor binding, Phylogenetic conservation and DNaseI hypersensitivity (DHS), have been conventionally used as generic markers for possible regulatory regions.

There are several mathematical algorithms that predict regulatory regions by interpreting data for one of these biological features. For instance, as part of the ENCODE Project the University of Washington (UW) and Duke University use DHS [12, 13], while the JASPER and

TRANSFAC databases use transcription-factor binding [10, 11]. A composite algorithm could be developed that utilizes several of the features together to provide a more elaborate description of regulatory elements across the genome. In this thesis, we started with the most basic tool i.e. visualizing these features across the length of a chromosome. When looking in a specific region, visual representation, besides being the simplest method of analysis, is often times better than most complex algorithms. Although excellent visualization tools such as the UCSC Genome Browser [14] exist, they are generic in nature and somewhat lack customizing ability and visual appeal. Here we describe a tool that focuses on highlighting regulatory regions in the genome or a part thereof with almost indefinite customizations.

2.2 Samples

2.2.1 Sample selection

2.2.1.1 *Biological markers for regulatory elements*

We selected the following biological features that indicate the presence of regulatory elements at specific locations and retrieved them from relevant public databases.

a) Since, some regulatory elements are known to be conserved across species due to their biological significance, conservation can be

used as a marker for regulatory elements. We chose the following properties indicating conservation.

i. PhastCons: Data for conservation across 46 vertebrate species was obtained from the UCSC genome browser database.

ii. Evolutionary Conserved Data (ECR): It provides conservation through pairwise alignment of genomes across species. We used the human alignment data with Dog, Mouse, Chicken and Zebrafish from NCBI Dcode database [15].

b) Transcription factor binding sites (TFBS): These are the sites where transcription factors bind at the start of the transcription process or at distal enhancers, and hence play a significant role in expression regulation. We used the public data for untreated samples from various labs participating in the ENCODE study. The CTCF, MEF2A and MEF2C transcription factors were considered for this study along with P300, a co-activator also indicative of the possible TFBS. ENCODE Tier 1 & Tier 2 cell lines from Stanford/Yale/USC/Harvard(SYDH) Universities and HudsonAlpha Institute of Biotechnology (HAIB) labs and all available cell lines for University of Texas-Austin(UTA) and University of Washington(UW) labs were used. Data from these cell lines were coalesced together.

c) DNaseI hypersensitive sites (DHS): These represent a measure of open chromatin and hence, act as a general marker for different

kinds of functional elements in the genome. ENCODE data for all available cell lines from University of Washington and Duke University were collected and coalesced together.

2.2.1.2 *Variants data from GWAS study of NOS1AP*

Location, effect size and frequency data for SNPs in the *NOS1AP* (NO-synthase adapter protein) gene locus of the human genome were obtained from a genome-wide analysis study of electro-cardiographic QT-interval performed in over 76,000 individuals of European ancestry (courtesy of Dr. Dan Arking). Additional common SNPs were obtained from the 1000 Genomes project.

To effectively study the locus, we also included tracks for recombination rate and genes. The genetic map of the human genome was retrieved from HapMap Phase II, release 22 [16]. It contains annotations of 3.1 million SNPs from several different human ancestry across the planet. Gene information from RefSeq database [17] was used for genes locations and structures.

2.2.2 Setting up the software

The software (Appendix A) is developed in R programming language [18] and requires the Rscript utility (comes with default R installation package). Data files for each track must be created in tab-delimited files

and placed in the same folder as the software. To run the software, a few basic parameters such as chromosomal location of the region of interest are needed and rest of the parameters depend on the changes made while customizing the software. A simple command line invocation might look like:

```
> Rscript final_plotter.R chr1:160290000-160310000
```

2.3 Results

2.3.1 Analysis of the *NOS1AP* locus

To demonstrate the software, we focused on the *NOS1AP* locus, whose effect on sudden cardiac death has been shown previously [19]. Data from all the above sources were plotted for 30kb region around *NOS1AP* locus on chromosome 1 as shown in figure 2.1. As can be seen in the figure, there are only a few significant SNPs that lie in regions with known DHS or TFBS. Overall, there appears to be a pattern of lower conservation at SNP locations. ECR values for dog, and to some extent mouse, which are present over a large portion of the human genome, are the only annotated conserved regions. Overall, apart from one well-studied sentinel SNP for QT-interval, rs12143842, we found no other SNPs that lie in annotated regions.

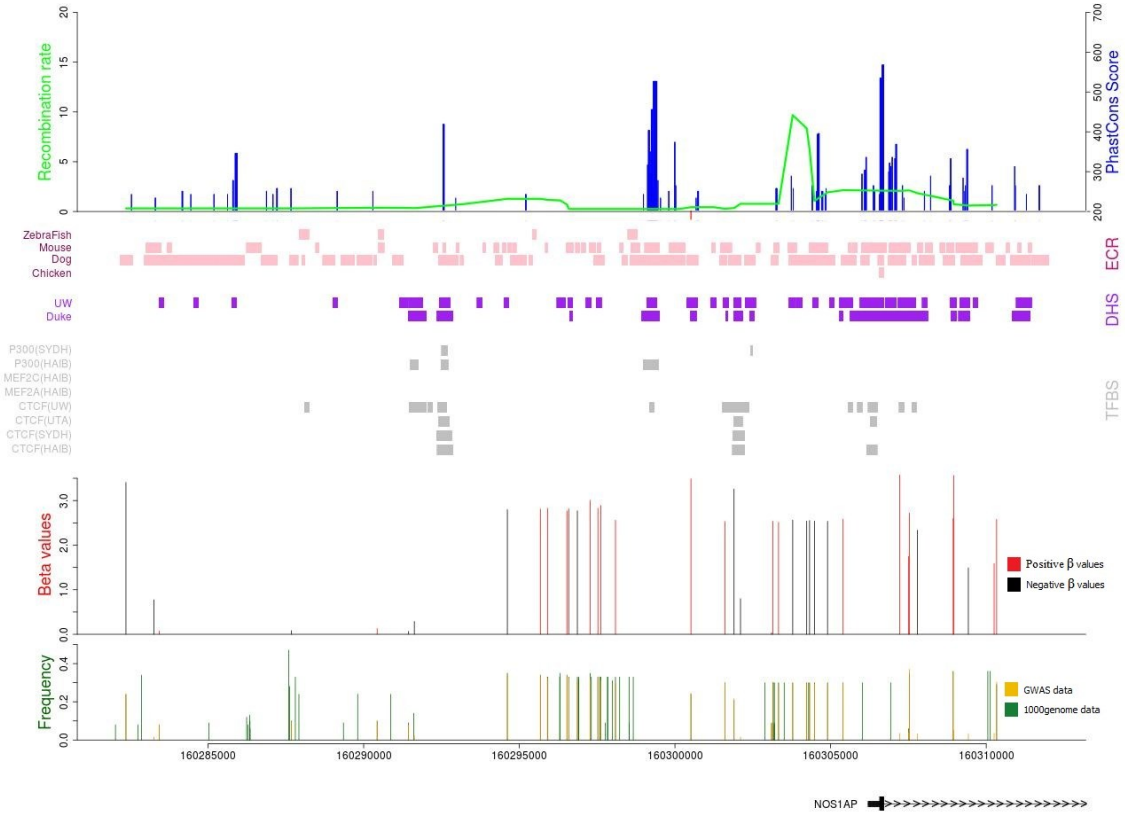

Figure 2.1 Software output for the 30kb region surrounding the NOS1AP locus, along the length of the chromosome on X-axis. (Top to bottom) Overlapping curves are recombination rates in green, and PhastCons scores in blue; ECR values for alignment with Human genome, transcription factor binding sites identified for different transcription factors (by labs in bracket); beta values represent the effect of SNPs in GWAS study of QT-interval (Positive being enhancing and negative being suppressing in effect); frequency of SNPs studies (GWAS SNPs in yellow and 1000 genomes imputed SNPs in green) and gene location and structure at the bottom.

2.3.2 Analysis of Beta-globin locus

We also used the software to briefly study the beta-globin locus. Figure 2.2 shows the plot of this locus. In this region, we observed inconsistencies among different data sets, and even for the same data types produced by different labs. For example, around position 5243000 on Chromosome 1, several sets of annotations are in agreement, however both P300 (done by SYDH lab) and CTCF (done by UT Austin) tracks don't show any signal, rather peaking at different location.

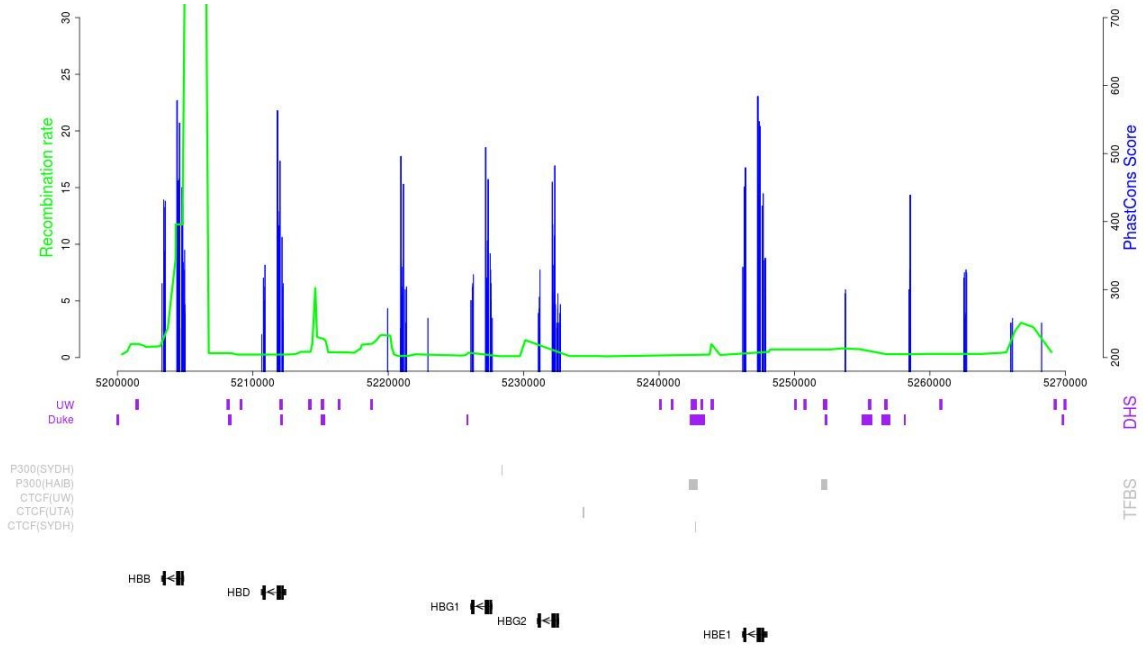

Figure 2.2 Software output for the 70kb region surrounding the beta-globin protein gene, along the length of the chromosome on X-axis. Overlapping at the top are recombination rates in green and PhastCons scores in blue, transcription factor binding sites identified for difference transcription factors (by labs in bracket), and gene location and structure.

2.4 Summary & Discussion

With the help of plotted results of two loci regions, we can see how this software can help researchers in visualizing their region of interest, study the available statistics and annotations, and overall, have a better understanding of the area under consideration. Ability to add tracks such as GWAS data, adjust range on y-axis and order tracks gives flexibility to the user. Although it has certain disadvantages compared to renowned tools such as the UCSC Genome browser, which can automatically fetch data for most tracks and provides better navigation and drag-and-drop features, our tool is simpler in its design and functionality and hence provides the user full control to customize visuals such as colors, type of plot for each track, overlapping tracks, etc. It also has an advantage in terms of exporting the generated charts to various image formats and PDF, which can be easily incorporated into documents.

On the other hand, close examination of the results of these two plots reveals several regions that are not annotated by one or more studies. This suggests the possibility that there might be other sites that are DNaseI sensitive or bound by transcription factors, but they are not strong enough to pass the threshold of the algorithms applied. This would also explain how algorithms tuned in a slightly different manner

might end up selecting few similar sites and many different regions to annotate. In order to validate our hypothesis, we decided to analyze the distribution of regulatory elements across the human genome, which is the topic for the second half of the thesis.

Chapter 3

Modelling DNaseI sensitivity across human genome

3.1 Introduction

Deoxyribonuclease I or DNase I is an enzyme that enables cutting of DNA sequence by breaking the chemical bond between adjacent nucleic acids. Under normal circumstances, the DNA in a eukaryotic cell is wrapped inside the nucleus by histone molecules in super-coiled state, known as chromatin. Chromatin is inaccessible to DNase, so even if DNase is added, virtually no reaction takes place. However, the chromatin opens during the transcription process to reveal parts of the DNA sequence to allow access to regulatory factors. DNase added in this system cuts the DNA at open chromatin positions. Hence, the sites that have excessive cutting by DNase, called DNaseI hypersensitive sites (DHS), are markers for accessible chromatin. As open chromatin is an indicator of underlying regulators of transcription process, DHS regions are considered generic markers for identification of different types of regulatory elements in the

genome and have been noted to correspond to promoters, enhancers, insulators, and other regulatory features [7, 8].

The Encylopedia of DNA Elements, or ENCODE, Project [11] has carried out genome-wide treatments with DNase across many cell lines. Public availability of this data allows us to study the distribution of DNase I sensitivity throughout the human genome, which effectively translates into analysis of functional parts of the genome which can then be used to identify and understand the causal SNPs in complex traits and diseases.

3.2 Samples and Methods

3.2.1 Sample selection and preparation

We used the alignment files provided by University of Washington (UW) as part of the ENCODE project. The files contain sequencing reads aligned to the human genome, which highlight DNA regions cut by DNase activity. Reads mapping to more than one location in the genome were removed, however, replicate reads were retained. We used the data unaltered. Data was collected for the following 7 cell lines (including replicates where available): cardiac fibroblasts (HCF), cardiac myocytes (HCM), embryonic stem cells (H1), undifferentiated embryonic stem cells (H7) and lymphoblast from different individuals (GM12864, GM12865, GM12878).

The entire human genome was split into 30bp bins. Each 36 bp read was then allocated to the bin where majority of its sequence lied. In case of a tie, random allocation was made to one of the tied bins. Once every read was allocated, number of reads in each bin was counted. This data, namely numbers of bins with specified number of reads, was then used to model the distribution.

At the extreme end, we see a small number of bins with up to thousands of reads that lie isolated to the distribution. When studied in detail, we found that most of these outliers belong to the same bin across cell lines. Since, this is unrealistic epigenetically, it is likely that these bins represent artifacts due to selective sequence advantage during DNA cutting, sequencing or other experimental procedures. Hence, we ignored bins with more than 250 reads per bin for the purpose of this study.

3.2.2 Model proposition and fitting

Under circumstances where the entire genome had equal sensitivity to DNase activity, the system could be modeled as a Poisson distribution with its mean equal to total number of reads divided by total number of bins and we could predict the number of bins with specified number of reads.

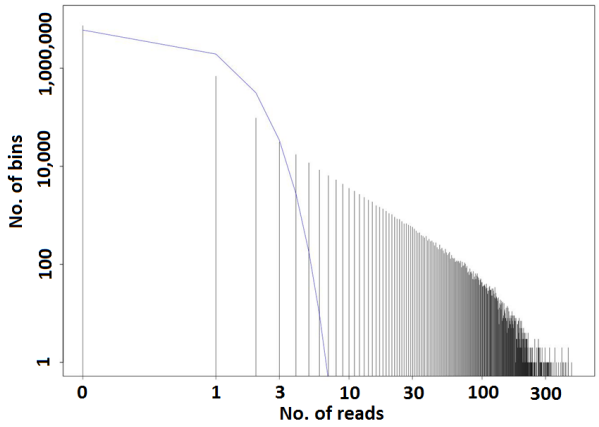

Figure 3.1 *Blue line shows ideal Poisson curve for uniformly sensitive DNA against bar curve of real values from chromosome 1 of HCF cell line (replicate 1) on a log-log plot.*

Figure 3.1 represents such a curve and highlights the fallacy in this argument, as we expect. Under a uniform distribution, no bin should contain more than 7 reads, but since some parts of the genome are highly sensitive, we see bins with number of reads greater than 100.

However, the smooth curve outlining the bar chart implies the existence of an intrinsic function that defines the distribution. We proposed that DNase sensitivity across the human genome follows a Gamma distribution. Choice of Gamma was based on two major criteria: its ability to take a variety of shapes based on its shape (r) and scale (a) parameters, and its conjugation with the Poisson distribution.

Hence, the distribution can be modeled as a Poisson distribution with its mean varying as a Gamma distribution with two parameters. Further complicating the model, in a competing process DNase cuts DNA

sequence at random locations. This process may be attributed to chromatin opening in some cells for base level transcription, DNA replication or other processes. Resultant reads align at insignificant regions, which were treated as noise and were modeled as a simple Poisson distribution. Hence, mathematically, our model can be represented as:

$P(k) = w\ \text{Poisson}(k; \lambda_g) + (1-w)\ \text{Poisson}(k; \lambda_r)$

where P(k) -> fraction of bins with k reads each

w -> fraction of reads following the Gamma distribution

$\lambda_g \sim \text{Gamma}(a,r)$

$\lambda_r \sim$ constant

We developed a script (available in Appendix B) to utilize the Maximum-likelihood estimation package in R that uses the quasi-Newton method to fit the data for individual chromosomes as well as for the entire genome for multiple cell lines.

3.3 Results

3.3.1 Parameters for final fitting

Fitting chromosome 1 data from the HCF cell line replicate 1 resulted in the following values for the three parameters of our model at the maximum likelihood:

a = 0.03448, r = 0.01629 and w = 0.49732

The resultant curve using these parameters gives us a better fit shown in figure 3.2. The list of parameters for individual chromosomes and the whole genome, from selected cell lines, is provided in Appendix C. It is interesting to observe that the value of w is always around 0.5,indicating that only about half the time DNase cuts are targeted based on sequence sensitivity, while about half the time cuts are random in nature.

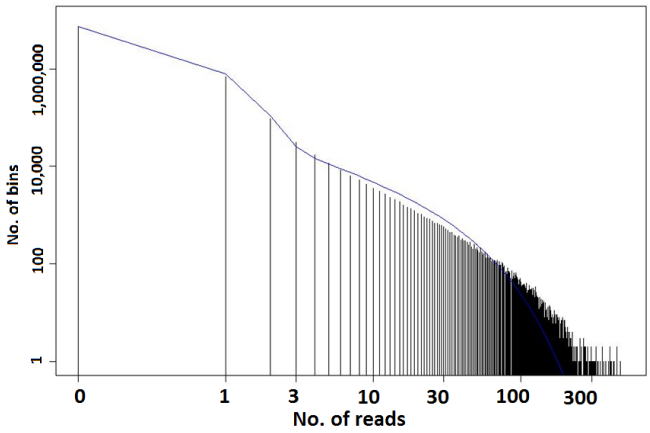

Figure 3.2 *Bar curve represents raw data from chromosome 1 of HCF cell line (replicate 1) on a log-log plot while blue line is the curve fitted using our model.*

3.3.2 Comparison of replicate datasets

We tried to perform a basic validation of our hypothesis by fitting the datasets for replicates, where available. We fitted individual chromosome data for both replicates for each of cell lines: HCF, HCM, H7, GM12865 and GM12878 and plotted the resulting parameters on two axes as represented in figure 3.3. Each point on the plot represents a parameter value estimated using data from one of the chromosomes from one of the cell lines.

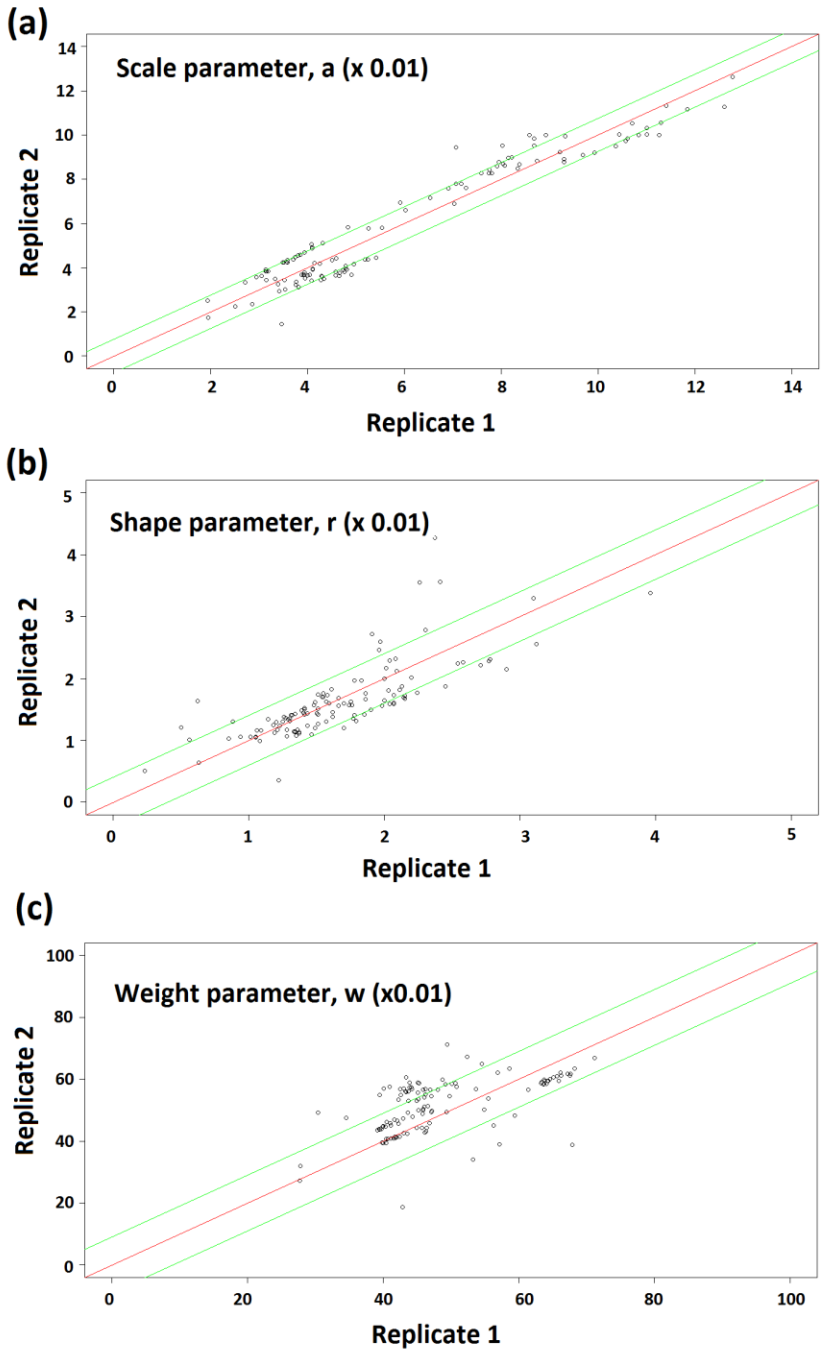

Figure 3.3 *Comparison of parameters between replicates. X-axis represents value of parameter in replicate 1 and Y-axis has its value in replicate 2. Red line is ideal situation, where the parameters are equal, and green lines are drawn at one standard deviation.*

Under ideal conditions, parameters from the two replicate would be equal and would lie on the red line. Although not on the line, observed parameters are very close to the ideal lines and most lie within single standard deviation. Since some of the deviation could be assigned to the experimental variations, we can infer that at the very least the model is not biased towards dataset and treats both replicates similarly.

3.3.3 Variation across chromosomes

Next, we compared the DNase sensitivity profiles of individual chromosomes within a cell line. Plots in figure 3.4 showGamma distributions for each chromosome for (a) HCF and (b) GM12864 cell lines, plotted using estimated parameters for best fit. At the left end, i.e. least DNase sensitive end, all the chromosomes are close together and are at their highest value, indicating that the majority of the genome is insensitive to DNase activity. As the levels of sensitivity increase, we observe gradually lesser parts of the genome being covered at those levels. Further, there is a sudden shift in the curve around DNase sensitivity value of 10, where curve falls much more steeply. This drop indicates that regions with more than 10 times the average sensitivity are much more rare. These regions can be classified as DHS sites with high confidence.

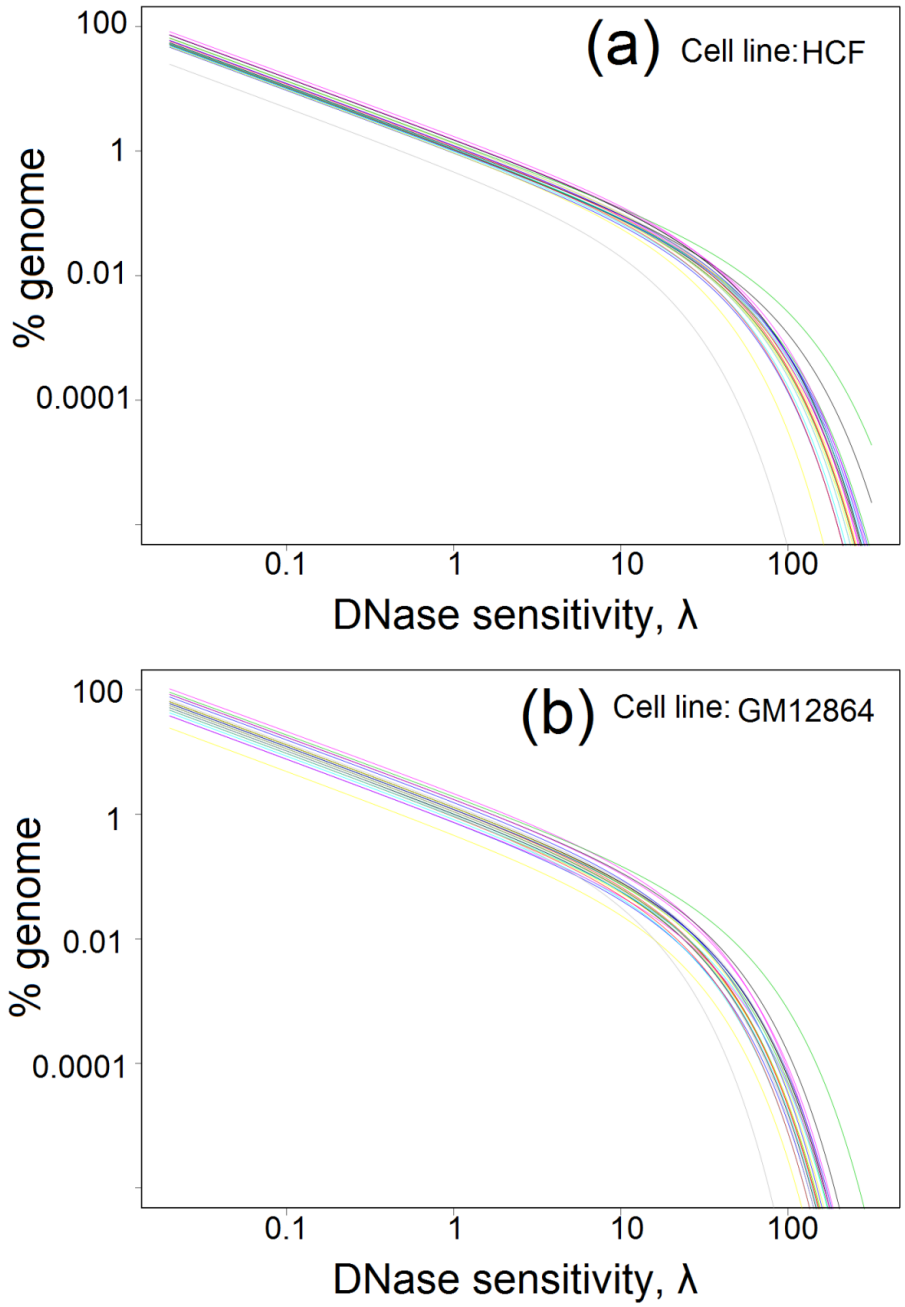

Figure 3.4 *Gamma distributions for each chromosome in (a) HCF cell line, (b) GM12864 cell line. Curve of each color is plotted for best-fit parameters for one chromosome. DNase sensitivity on X-axis represents number of reads that would ideally align to that region for average genome coverage of one and Y-axis represents fraction of genome with that coverage.*

We expect the DHS sites, shown at the tight-most end on the plot, to be related to gene density and gene coverage. If we look at the curves in figure 3.4, the highest curves belong to Chromosomes 19 (green) and 17 (black) which have the largest numbers of genes and maximum gene coverage per base-pair among all chromosomes. The lowest curves correspond to chromosomes Y (grey) and X (yellow) which have minimum gene coverage and are among the chromosomes with least number of genes per base-pair. This was observed among other cell lines (except for missing Y chromosome for cell lines obtained from females). Therefore, generally, gene density and gene coverage seem to be directly correlated to the fraction of DHS sites in the region.

3.3.4 Differences among different cell lines

From figure 3.4 and similar curves from other cell lines, we also notice that the left part of the curve is similar among cell lines whereas the right part of the curve drops at different rates. This difference in cell line parameters is more pronounced when considering individual parameters estimated by fitting the genome-wide data (figure 3.5). Only the HCF and HCM cell lines, which have close biological relationship have comparable parameters. An important observation is the significant difference among GM cell lines, all of which originated from lymphobastoids, although from different individuals.

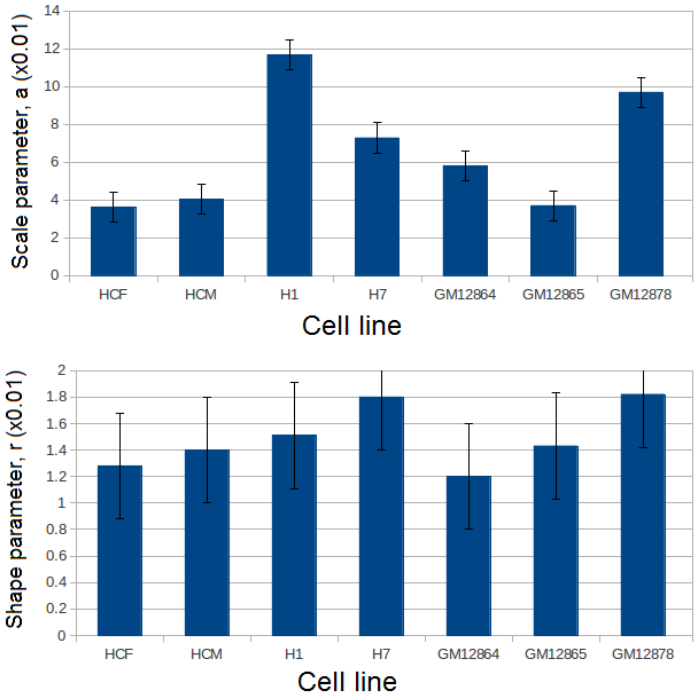

Figure 3.5 *Comparison of genome-wide fitting parameters from different cell lines. Error bars on each side represent one standard deviation difference (calculated in comparison of replicates).*

3.4 Conclusion & Discussion

The better fit of the model over several cell lines in this chapter support the fact that underlying sensitivity distribution of the human genome could be modeled as a Gamma distribution. This conclusion is further bolstered by the study of replicates, which showed that parameters for fitting the replicate data are within the limits of experimental errors. Moreover, when comparing different chromosomes from same cell line, we see a correlation between DHS and gene density and coverage, as expected.

From these results, we can also understand the following about the distribution of DNase sensitivity across the human genome. Value of w (i.e. fraction of reads following gamma distribution) is close to 0.5, which means that a large part of the genome likely does not participate in regulation at all. Of the remaining portion, a major portion (shown in the left part of gamma curves) has very low sensitivity. And only a very small portion (shown in the right part of gamma curves) is truly DNase I hypersensitive.

Although there are potentially some data artifacts stemming from the filters used on the UW data, such as not removing replicate reads, consistency is observed in overall shape of the curves, even when replicates are removed, as well as when using data from Duke University. Further, the method could be applied to find parameters for future data, as it becomes available. Also, variations in the algorithm such as using 1 kb bin instead of 30 bp ones, or binning on the basis of 5' end of the read rather than majority binning, do not alter the shape of the curve and have minimal effects over final parameters.

In an extension of this study, we are studying the parameters in different types of elements in the genome such as introns, untranslated regions, repeats etc. One possible next step could be to study the similar distribution for transcription factor binding sites, PhastCons or other

features. In the long run, these distributions could be used to assign a score for each feature which can then be combined to give overall likelihood of a region being regulatory or otherwise.

References

1. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Trans. Royal Soc. Edin. 52, 399-433, 1918.

2. Manolio TA, Collins FS, et al. Finding the missing heritability of complex diseases. Nature 461:747-753, 2009.

3. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 273:1516-1517, 1996.

4. Collins F, Guyer M, Chakravarti A. Variations on a theme: cataloging human DNA sequence variation. Science 278:1580-1581, 1997.

5. Hindorff LA, Sethupathy P, et al.: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA 106:9362-9367, 2009;

6. Schlesinger J, Schueler M, et al. The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. PLoS Genet. 7:e1001313, 2011.

7. Gross DS, Garrard WT. Nuclease hypersensitive sites in chromatin. Annual review of biochemistry. 57:159-97, 1988.

8. Stalder J, Larsen A, Engel JD, et al. Tissue-specific DNA cleavages

in the globin chromatin domain introduced by DNAase I. Cell. 20(2):451-60, 1988.

9. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 32(Database issue):D91-4, 2004.

10. TRANSFAC: an integrated system for gene expression regulation. Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruess, M., Reuter, I., Schacherer, F.Nucleic Acids Res. 28:316-319, 2000.

11. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 306(5696):636-40, 2004.

12. Sabo, Peter J., et al. "Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries." Proceedings of the National Academy of Sciences of the United States of America 101.13 (2004): 4537-4542.

13. Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics. 24(21):2537-8, 2008.

14. Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. Curr Protoc Bioinformatics. Chapter 1:Unit1.4, 2012.

15. G.G. Loots and I. Ovcharenko. ECRbase: Database of Evolutionary Conserved Regions, Promoters, and Transcription Factor Binding Sites in Vertebrate Genomes, Bioinformatics. 23(1):122-4, 2007.

16. Thorisson, G.A., Smith, A.V., Krishnan, L., and Stein, L.D. The International HapMap Project Web site. Genome Research,15:1591-1593, 2005.

17. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 40(Database issue):D130-5, 2012.

18. Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5(3):299-314, 1996.

19. Arking DE, Pfeufer A, et al. A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. Nat. Genet. 38:644-651, 2006.

Appendices

Appendix A

Code for Annotation Visualization Software

```
# Reading command-line arguments

args<-commandArgs(TRUE)

chr <- as.integer(unlist(strsplit(args[9],":|-"))[1])

xstart <- as.integer(unlist(strsplit(args[9],":|-"))[2])

xend <- as.integer(unlist(strsplit(args[9],":|-"))[3])

phastCons_cutoff<-c(as.numeric(args[3]),as.numeric(args[4]))

recomb_cutoff<-c(as.integer(args[5]),as.integer(args[6]))

freq_cutoff<-c(as.numeric(args[1]),as.numeric(args[2]))

zs_cutoff<-c(as.numeric(args[7]),as.numeric(args[8]))


# Reading data files

loci <- read.csv("Dan_data/loci.csv")

beta <- read.csv("Dan_data/beta_l3.csv")

phastCons <-

read.csv("phast_cons_data/phast_cons_el_vertebrate_l3.csv")

gene <- read.csv("gene_l3.csv")

all_snps <- read.csv("1000_genome_data/all_snp_freq_l3.csv")
```

```
ecr_data <- read.csv("ECR_data/l3.csv")

dhs_data <- read.csv("encode_DHS_data/hcf_hcm_l3.csv")

tfbs_data <- read.csv("encode_TFBS_data/hcf_hcm_l3.csv")

read_data <- read.csv("read_count.csv")


# Selecting data for region of interest

loci <- loci[loci$Position>xstart & loci$Position<xend &

loci$Chromosome==chr,]

beta <- beta[beta$Position>xstart & beta$Position<xend,]

phastCons <- phastCons[phastCons$chromEnd>xstart &

phastCons$chromStart<xend,]

gene <- gene[gene$txEnd>xstart & gene$txStart<xend,]

all_snps <- all_snps[all_snps$Position>xstart & all_snps$Position<xend,]

ecr_data <- ecr_data[ecr_data$End>xstart & ecr_data$Start<xend,]

dhs_data <- dhs_data[dhs_data$end>xstart & dhs_data$start<xend,]

tfbs_data <- tfbs_data[tfbs_data$end>xstart & tfbs_data$start<xend,]

read_data <- read_data[read_data$end>xstart & read_data$start<xend,]

chr_map=read.table(paste("genetic_maps/genetic_map_l3.txt",sep=""),hea

der=T,sep=" ")

chr_map <- chr_map[chr_map$position>xstart &

chr_map$position<xend,]

spacing <- (xend-xstart)/5000
```

```
#Setting up charting configurations
jpeg(paste("Chromosome",chr,".jpg",sep=""),height=960,
width=1280,units="px")
layout(matrix(1:8,8,1),heights=c(20,15,2*length(levels(ecr_data$Species)),
2*length(levels(dhs_data$source)),2*length(levels(dhs_data$source)),15,1
2,12))
# Plotting PhastCons data
par(mar=c(0,10,1,5))
phast_cons_pos <- c()
phast_cons_score <- c()
if (length(phastCons[,1]) > 0)
for (j1 in 1:length(phastCons[,1]))
{
cons_pos <-
seq(phastCons$chromStart[j1],phastCons$chromEnd[j1],by=spacing)
cons_pos <- c(cons_pos,phastCons$chromEnd[j1])
phast_cons_pos <- c(phast_cons_pos,cons_pos)
phast_cons_score <-
c(phast_cons_score,rep(phastCons$score[j1],length(cons_pos)))
}
plot(phast_cons_pos,phast_cons_score,type="h",col="blue",axes=F,ann=F,
xlim=c(xstart,xend),ylim=phastCons_cutoff,cex.axis=2,cex.lab=2)
```

```
axis(side=4,cex.axis=1.5)
mtext("PhastCons Score",side=4,col="blue",line=3,cex=1.5)
abline(h=phastCons_cutoff[1]*1.045-
0.045*phastCons_cutoff[2],lwd=30,col="white")
rug(loci$Position,col="red",quiet=T,lwd=1.5,ticksize=0.04)

#Plotting recombination rates
par(new=T)
plot(chr_map$position,chr_map$COMBINED_rate,type="l",col="green",lw
d=3,main=paste("Chromosome",chr),axes=F,ann=F,xlim=c(xstart,xend),yli
m=recomb_cutoff,cex.axis=2,cex.lab=2,cex.main=3)
axis(side=2,cex.axis=1.5)
mtext("Recombination rate",side=2,col="green",line=3,cex=1.5)
abline(h=recomb_cutoff[1])

#Plotting ECR values
specie_count <- 1
for (specie in levels(ecr_data$Species))
{
        positions <- c()
        ecr_starts <- ecr_data$Start[ecr_data$Species==specie]
        ecr_ends <- ecr_data$End[ecr_data$Species==specie]
```

```
        if (length(ecr_starts) > 0)
        for (l1 in 1:length(ecr_starts))
        {
                positions <-
c(positions,seq(ecr_starts[l1],ecr_ends[l1],by=spacing))
                positions <- c(positions,ecr_ends[l1])
        }
        if (specie_count != 1) par(new=T)
        plot(positions,matrix(specie_count,length(positions),1),axes=F,col="
pink",pch="|",ann=F,xlim=c(xstart,xend),ylim=c(0,5),cex=1.5)
        text(xstart,specie_count,paste(specie,"
"),adj=1,xpd=T,cex=1.5,col="deeppink4")
        specie_count <- specie_count+1
}
mtext("ECR",side=4,col="deeppink3",line=3,cex=1.5)

#Plotting DHS sites
source_count <- 1
for (data_source in levels(dhs_data$source))
{
        positions <- c()
        dhs_starts <- dhs_data$start[dhs_data$source==data_source]
```

```
        dhs_ends <- dhs_data$end[dhs_data$source==data_source]
        if (length(dhs_starts) > 0)
        for (l1 in 1:length(dhs_starts))
        {
                positions <-
c(positions,seq(dhs_starts[l1],dhs_ends[l1],by=spacing))
                positions <- c(positions,dhs_ends[l1])
        }
        if (source_count != 1) par(new=T)
        plot(positions,matrix(source_count,length(positions),1),axes=F,col="
purple",pch="|",ann=F,xlim=c(xstart,xend),ylim=c(0,length(levels(dhs_dat
a$source))+1),cex=1.5)
        text(xstart,source_count,paste(data_source,"
"),adj=1,xpd=T,cex=1.5,col="purple")
        source_count <- source_count+1
}
mtext("DHS",side=4,col="purple",line=3,cex=1.5)

#Plotting TFBS regions
source_count <- 1
for (data_source in levels(tfbs_data$source))
{
```

```
        positions <- c()
        tfbs_starts <- tfbs_data$start[tfbs_data$source==data_source]
        tfbs_ends <- tfbs_data$end[tfbs_data$source==data_source]
        if (length(tfbs_starts) > 0)
        for (l1 in 1:length(tfbs_starts))
        {
                positions <-
c(positions,seq(tfbs_starts[l1],tfbs_ends[l1],by=spacing))
                positions <- c(positions,tfbs_ends[l1])
        }
        if (source_count != 1) par(new=T)
        plot(positions,matrix(source_count,length(positions),1),axes=F,col="
gray75",pch="|",ann=F,xlim=c(xstart,xend),ylim=c(0,length(levels(tfbs_dat
a$source))+1),cex=1.5)
        text(xstart,source_count,paste(data_source,"
"),adj=1,xpd=T,cex=1.5,col="grey75")
        source_count <- source_count+1
}

mtext("TFBS",side=4,col="grey75",line=3,cex=1.5)
```

```
#Plotting SNPs effect size
plot(beta$Position[beta$Beta>0],beta$Beta[beta$Beta>0],type="h",col="re
d",ann=F,axes=F,xlim=c(xstart,xend),ylim=c(0,max(c(0,abs(beta$Beta)))))
lines(beta$Position[beta$Beta<0],-
beta$Beta[beta$Beta<0],type="h",col="black")
mtext("Beta values",side=2,col="red",line=3,cex=1.5)
abline(h=0)
axis(side=2,cex.axis=1.5)

#Plotting SNPs frequencies
plot(all_snps$Position,all_snps$Frequency,type="h",col="darkgreen",ann=
F,axes=F,xlim=c(xstart,xend),ylim=freq_cutoff)
lines(beta$Position,beta$Frequency,type="h",col="darkorange")
mtext("Frequency",side=2,col="darkgreen",line=3,cex=1.5)
abline(h=freq_cutoff[1])
axis(side=1,cex.axis=1.5)
axis(side=2,cex.axis=1.5)

#Plotting gene structure
exon_starts = lapply(strsplit(as.matrix(gene$exonStarts),","),as.numeric)
exon_ends = lapply(strsplit(as.matrix(gene$exonEnds),","),as.numeric)
cds_start = lapply(as.matrix(gene$cdsStart),as.numeric)
```

```
cds_end = lapply(as.matrix(gene$cdsEnd),as.numeric)
intron_dist <- (xend-xstart)/100
par(mar=c(0,10,2,5))
if (length(gene[,1])>0)
for (k1 in 1:length(gene[,1]))
{
        exons<-c()
        introns_pos<-c()
        introns_neg<-c()
        UTRs<-seq(exon_starts[[k1]][1],cds_start[[k1]],by=spacing)
        UTRs<-
c(UTRs,seq(cds_end[[k1]],exon_ends[[k1]][length(exon_ends[[k1]])],by=spa
cing),exon_ends[[k1]][length(exon_ends[[k1]])])
        for (k2 in 1:length(exon_starts[[k1]]))
        {
                ex_str <- max(cds_start[[k1]],exon_starts[[k1]][k2])
                ex_end <- min(cds_end[[k1]],exon_ends[[k1]][k2])
                exons <- c(exons,seq(ex_str,ex_end,by=spacing))
                exons <- c(exons,ex_end)
                if (k2 != 1)
                {
                        gap <- exon_starts[[k1]][k2]-exon_ends[[k1]][k2-1]
```

```
                    if (gap > 1.0*intron_dist)
                    {
                        spec_intron_dist <- intron_dist
                        spec_intron_dist <- gap/round(gap/intron_dist)
                        intron_region <- seq(exon_ends[[k1]][k2-
1]+0.5*spec_intron_dist,exon_starts[[k1]][k2]-
0.5*spec_intron_dist,by=spec_intron_dist)
                    }
                    else    intron_region <- c()
                    if (gene$strand[k1]=="+") introns_pos <-
c(introns_pos,intron_region)
                    else  introns_neg <- c(introns_neg,intron_region)
                }
        }
        plot(UTRs,matrix(k1,length(UTRs),1),pch="|",cex=1,axes=F,ann=F,
xlim=c(xstart,xend),ylim=c(length(gene[,1])+1,-0))
        points(introns_pos,matrix(k1,length(introns_pos),1),pch=">",cex=2)
        points(introns_neg,matrix(k1,length(introns_neg),1),pch="<",cex=2)
        points(exons,matrix(k1,length(exons),1),pch="|",cex=2)
        segments(gene$txStart[k1],k1,gene$txEnd[k1],k1)
        if (as.numeric(gene$txStart[k1])<xstart &
as.numeric(gene$txEnd[k1])>xstart)
```

```
                text(xstart,k1,paste(gene$name[k1],"
"),cex=1.5,adj=1,xpd=T)
        else
                text(gene$txStart[k1],k1,paste(gene$name[k1],"
"),cex=1.5,adj=1,xpd=T)
        par(new=T)
}
garbage <- dev.off()
```

Appendix B

Code for Maximum-likelihood fitting of the model

```
library("stats4")
setwd("~/Dropbox/labwork/HCF_rep1")  # Location to cell line data
bin_size <- 30


# Functions to calculate likelihood for given set of parameters
p_k_factor <- function(a,r,k) { (r+k)/((1+k)*(1+a)) }
eff_p <- function(pr) { log(pr/sum(pr)) }
ll <- function(a,r,w){
  if (a>=0 && r>=0 && w<=1 && w>=0){
    p <- w*(a/(1+a))^r
    for (k in 1:max(rng)) p[k+1] = p[k] * p_k_factor(a,r,k-1)
    lr<-(m-w*r/a)/(1-w)
    if(lr > 0)
    {
      p <- p+(1-w)*dpois(0:max(rng),lr)
      -sum(bin_counts[rng]*eff_p(p[rng]))
    }else Inf
  }else Inf
}
```

```
# Number of unknown nucleotides (N) in each chromosome
Ns <-
c(23970000,4994851,3225294,3492600,3220000,37200020,3785000,34
75100,21070000,4220005,3877000,3370501,19580000,381201,208366
23,11470000,3400000,3420015,3320000,3520000,13023203,16410004
,4170000,33720000)
# Looping for chromosomes 1 to 22, X &Y
for (chr_num in 1:24){
if(chr_num==23){
  chr <- 'X'
} else {
  if(chr_num==24) chr <- 'Y'
  else  chr <- chr_num
}
# Calculating number of bins with each number of reads value.
bin_counts <-
read.csv(paste("bincounter_chr",chr,"_read_locations_rep1.txt",sep=""))[,1]
rng <- 1:min(250,length(bin_counts))
N_bins <- Ns[chr_num]/bin_size
bin_counts[1] <- bin_counts[1]-N_bins
# Calculating mean coverage of reads
m <- sum((rng-1)*bin_counts[rng])/sum(bin_counts)
```

```
# Calling mle function to fit the model
# Initial parameters are slightly altered in case of error with default initial
parameters
r_start<-m/2
o<-0
while(is.numeric(o) & r_start < 3*m/4)
{
  o<-
tryCatch(mle(ll,start=list(a=1,r=r_start,w=0.5)),error=function(e){return(0)}
)
  r_start <- r_start + 0.01
}
while(is.numeric(o) & r_start > m/4)
{
  o<-
tryCatch(mle(ll,start=list(a=1,r=r_start,w=0.5)),error=function(e){return(0)}
)
  r_start <- r_start - 0.01
}
if(is.numeric(o)) print(paste(chr))
else print(paste(o@coef))
}
```

Appendix C

List of estimated parameters for fitting whole genome in various cell lines

Cellline	**a** (x0.01)	**r** (x0.01)	**w** (x0.01)
HCF	3.63	1.28	51.0
HCM	4.04	1.40	51.9
H1	11.67	1.51	51.7
H7	7.28	1.80	63.2
GM12864	5.81	1.20	46.1
GM12865	3.70	1.43	42.5
GM12878	9.69	1.82	43.7
Th1	3.02	0.96	55.6
Th2	4.93	1.62	52.3

List of estimated parameters for fitting individual chromosomes in various cell lines

HCF Cell line

	Replicate 1			Replicate 2		
	a (x0.01)	r (x0.01)	w (x0.01)	a (x0.01)	r (x0.01)	w (x0.01)
Chr1	3.45	1.63	49.73	3.52	1.57	47.16
Chr2	3.64	1.13	56.89	4.00	1.37	46.18
Chr3	3.69	1.14	56.59	3.93	1.34	45.86
Chr4	4.33	0.99	52.97	4.51	1.08	44.86
Chr5	3.94	1.17	54.22	4.11	1.28	46.03
Chr6	3.68	1.87	47.57	3.91	2.45	34.51
Chr7	3.78	1.12	54.60	3.93	1.19	47.15
Chr8	3.93	1.17	55.58	4.10	1.36	45.13
Chr9	3.37	1.14	58.68	3.77	1.33	50.62
Chr10	3.69	1.27	56.61	4.04	1.51	46.85
Chr11	2.94	1.11	58.55	3.42	1.37	50.08
Chr12	3.01	1.08	58.39	3.53	1.35	49.19
Chr13	4.18	1.05	51.27	4.25	1.05	46.65
Chr14	3.49	1.30	45.11	3.33	0.88	56.29
Chr15	3.67	1.42	57.60	3.87	1.51	50.80
Chr16	3.43	1.43	51.19	3.15	1.34	45.99
Chr17	2.36	1.20	67.25	2.86	1.70	52.41
Chr18	4.41	1.13	53.38	4.59	1.34	42.15
Chr19	1.75	1.46	62.10	1.95	1.62	56.87
Chr20	3.22	1.31	55.00	3.76	1.79	42.64
Chr21	3.25	1.15	53.99	3.46	1.22	49.43
Chr22	3.24	1.66	55.07	3.39	1.86	46.20
ChrX	5.81	1.01	38.97	5.54	0.56	57.11
ChrY	9.45	0.51	38.66	7.06	0.23	67.91

HCM Cell line

	Replicate 1			Replicate 2		
	a (x0.01)	r (x0.01)	w (x0.01)	a (x0.01)	r (x0.01)	w (x0.01)
Chr1	3.61	1.83	53.63	4.30	1.61	45.16
Chr2	4.08	1.50	56.86	4.79	1.41	44.18
Chr3	3.82	1.42	55.75	4.76	1.41	43.36
Chr4	4.46	1.16	57.27	5.42	1.09	43.94
Chr5	3.98	1.35	57.30	4.79	1.30	44.09
Chr6	3.91	2.28	49.22	4.81	2.77	30.36
Chr7	3.63	1.17	58.77	4.65	1.21	45.33
Chr8	4.16	1.44	57.03	4.95	1.43	43.00
Chr9	3.77	1.45	59.90	4.66	1.40	48.74
Chr10	3.67	1.38	60.60	4.91	1.62	43.42
Chr11	3.50	1.60	56.09	4.33	1.59	43.72
Chr12	3.66	1.57	56.95	4.58	1.74	40.09
Chr13	4.36	1.34	52.98	5.18	1.14	43.73
Chr14	3.52	1.03	64.86	3.95	0.85	54.55
Chr15	3.88	1.68	58.90	4.74	1.66	45.12
Chr16	3.40	1.44	56.08	4.08	1.50	43.33
Chr17	3.13	2.02	58.87	3.82	2.20	43.94
Chr18	4.38	1.31	56.93	5.25	1.30	42.47
Chr19	2.25	2.46	50.00	2.50	1.96	45.89
Chr20	3.82	1.81	54.90	4.58	2.03	39.47
Chr21	3.63	1.38	56.61	4.29	1.26	48.01
Chr22	3.43	1.87	57.60	4.28	2.13	40.88
ChrX	6.89	1.64	34.11	7.03	0.62	53.18
ChrY	9.52	1.21	18.66	8.02	0.50	42.84

H1 Cell line

	Replicate 1		
	a (x0.01)	r (x0.01)	w (x0.01)
Chr1	11.04	1.77	51.35
Chr2	12.79	1.41	51.12
Chr3	12.92	1.40	50.80
Chr4	14.72	1.15	53.41
Chr5	13.77	1.43	52.02
Chr6	12.43	2.49	40.40
Chr7	12.35	1.39	52.79
Chr8	15.25	1.59	50.71
Chr9	12.64	1.47	54.33
Chr10	12.94	1.64	50.38
Chr11	11.13	1.70	50.88
Chr12	11.37	1.58	50.44
Chr13	15.90	1.28	52.30
Chr14	11.31	1.03	59.72
Chr15	12.65	1.76	51.02
Chr16	9.85	1.94	48.65
Chr17	9.89	2.57	48.04
Chr18	13.92	1.31	50.32
Chr19	6.48	3.15	47.68
Chr20	12.13	2.36	44.50
Chr21	9.91	1.04	53.02
Chr22	11.34	2.75	44.42
ChrX	17.59	0.71	60.99
ChrY	Female sample		

H7 Cell line

	Replicate 1			Replicate 2		
	a (x0.01)	r (x0.01)	w (x0.01)	a (x0.01)	r (x0.01)	w (x0.01)
Chr1	5.26	2.78	61.44	5.78	2.31	56.64
Chr2	8.22	2.07	63.77	8.99	1.61	59.39
Chr3	8.07	2.07	63.64	8.63	1.59	58.33
Chr4	8.68	1.77	63.68	9.51	1.35	59.84
Chr5	8.14	1.98	64.28	8.95	1.56	59.54
Chr6	7.75	3.12	54.94	8.44	2.56	50.08
Chr7	7.80	1.90	65.61	8.27	1.49	60.94
Chr8	8.68	2.14	63.33	9.84	1.69	58.77
Chr9	8.04	2.04	66.12	8.69	1.59	62.17
Chr10	7.95	2.15	64.59	8.78	1.72	60.02
Chr11	7.17	2.24	65.12	7.79	1.77	60.61
Chr12	7.27	2.15	64.16	7.60	1.67	59.50
Chr13	8.92	1.85	63.41	9.99	1.42	59.13
Chr14	7.07	1.49	71.18	7.80	1.20	66.93
Chr15	7.59	2.07	67.58	8.26	1.73	61.81
Chr16	6.53	2.11	67.47	7.16	1.82	61.21
Chr17	6.02	2.54	67.16	6.59	2.24	61.76
Chr18	8.58	2.00	63.71	9.99	1.65	58.50
Chr19	4.14	2.90	58.65	4.20	2.15	63.53
Chr20	7.74	2.71	64.07	8.27	2.21	58.55
Chr21	7.91	1.78	66.20	8.58	1.41	61.09
Chr22	6.91	2.58	65.86	7.58	2.26	59.54
ChrX	8.37	1.57	68.18	8.66	1.30	63.42
ChrY	Female sample					

GM12864 Cell line

	Replicate 1		
	a (x0.01)	r (x0.01)	w (x0.01)
Chr1	5.51	1.43	45.12
Chr2	6.36	1.18	44.02
Chr3	6.28	1.15	43.99
Chr4	6.64	0.81	43.79
Chr5	6.50	1.09	44.44
Chr6	5.81	2.39	32.22
Chr7	6.10	1.02	46.27
Chr8	6.87	1.19	43.28
Chr9	6.62	1.11	48.99
Chr10	6.47	1.29	44.50
Chr11	5.63	1.25	44.95
Chr12	5.26	1.27	45.86
Chr13	7.05	0.91	45.38
Chr14	5.16	0.82	56.79
Chr15	6.04	1.37	47.68
Chr16	5.50	1.48	48.21
Chr17	4.91	1.87	48.82
Chr18	7.19	1.03	42.63
Chr19	3.44	2.09	52.09
Chr20	6.31	1.72	42.77
Chr21	5.48	0.98	50.83
Chr22	5.43	1.90	46.69
ChrX	7.46	0.47	56.47
ChrY	12.48	1.09	27.24

GM12865 Cell line

	Replicate 1			Replicate 2		
	a (x0.01)	r (x0.01)	w (x0.01)	a (x0.01)	r (x0.01)	w (x0.01)
Chr1	3.15	1.54	43.50	3.85	1.70	42.42
Chr2	3.74	1.32	41.06	4.47	1.41	40.78
Chr3	3.85	1.39	40.68	4.59	1.48	40.58
Chr4	4.09	1.01	40.40	4.88	1.06	40.68
Chr5	3.93	1.27	41.98	4.68	1.36	41.30
Chr6	3.58	3.10	27.73	4.35	3.30	27.34
Chr7	3.51	1.18	43.01	4.23	1.25	42.48
Chr8	4.09	1.31	39.97	4.93	1.41	39.40
Chr9	3.70	1.20	46.76	4.37	1.29	45.80
Chr10	3.80	1.48	41.72	4.55	1.57	41.40
Chr11	3.18	1.41	41.83	3.83	1.52	41.02
Chr12	3.14	1.49	42.36	3.83	1.62	41.40
Chr13	4.08	1.06	41.52	5.06	1.16	40.93
Chr14	3.14	0.94	55.51	3.92	1.06	53.76
Chr15	3.49	1.58	45.68	4.23	1.73	44.25
Chr16	2.94	1.55	46.32	3.57	1.76	43.07
Chr17	2.71	2.08	46.38	3.33	2.32	44.17
Chr18	4.31	1.25	39.76	5.11	1.30	39.61
Chr19	1.94	1.97	59.41	2.51	2.59	48.31
Chr20	3.59	1.83	40.39	4.24	1.97	39.32
Chr21	3.57	1.29	44.91	4.27	1.34	44.23
Chr22	3.05	2.04	46.15	3.62	2.29	42.75
ChrX	4.83	0.63	49.30	5.84	0.64	49.42
ChrY	Female sample					

GM12878 Cell line

	Replicate 1			Replicate 2		
	a (x0.01)	r (x0.01)	w (x0.01)	a (x0.01)	r (x0.01)	w (x0.01)
Chr1	9.20	2.17	47.27	9.92	2.01	42.88
Chr2	10.02	1.63	44.99	11.01	1.75	41.07
Chr3	10.01	1.57	44.57	11.26	1.76	40.34
Chr4	11.17	1.07	44.61	11.84	1.28	39.86
Chr5	10.55	1.56	45.81	11.30	1.66	41.01
Chr6	9.49	3.38	31.91	10.36	3.96	27.78
Chr7	9.72	1.51	46.86	10.57	1.51	41.57
Chr8	11.32	1.60	43.82	11.40	1.70	39.49
Chr9	10.33	1.70	49.94	11.01	1.55	45.21
Chr10	9.83	1.76	45.72	10.61	1.86	42.22
Chr11	8.91	1.97	46.13	9.29	1.78	40.50
Chr12	9.09	2.00	46.59	9.68	2.00	42.07
Chr13	11.28	1.10	43.88	12.60	1.46	39.66
Chr14	8.77	1.24	56.90	9.29	1.23	53.66
Chr15	10.01	2.12	47.79	10.84	2.09	44.32
Chr16	8.83	2.72	49.24	8.74	1.91	43.62
Chr17	8.49	3.56	50.25	8.34	2.41	45.92
Chr18	10.02	1.24	44.04	10.44	1.43	39.41
Chr19	6.94	4.27	54.49	5.91	2.37	49.79
Chr20	10.54	2.78	43.47	10.70	2.30	39.07
Chr21	9.94	1.74	48.95	9.32	1.51	45.78
Chr22	9.24	3.55	49.42	9.21	2.26	47.14
ChrX	12.63	1.06	44.86	12.78	1.05	40.02
ChrY	Female sample					

Curriculum Vitae

Vatsal Agarwal

EDUCATION

- Johns Hopkins University, Baltimore, MD, USA (2011-13)
  Masters of Science in Engineering, Biomedical Engineering

- Indian Institute of Technology, Roorkee, India (2005-09)
  Bachelors of Technology, Biotechnology

RESEARCH EXPERIENCE

- Johns Hopkins Medical Institute, Baltimore, MD, USA
  Graduate Research Assistant (2012-2013)
  Supervisor: Dr. Aravinda Chakravarti
  Project: Study of markers for regulatory elements in Human genome.

- Indian Institute of Technology, Roorkee, India
  Undergraduate project (2008-2009)
  Supervisor: Dr. Ritu Barthwal
  Project: Method to verify and refine structure of biomolecules, obtained using NMR machines.

- Ludwig Maximilians University, Gene Center, Munich, Germany
  Summer student (2008)
  Supervisor: Dr. Johannes Söding
  Project: PDBalert: automatic, recurrent remote homology tracking and protein structure prediction

- Indian Institute of Technology, Kanpur, India
  Undergraduate researcher (2007)
  Supervisor: Dr. Ramasubbu Sankararamakrishnan
  Project: MIPModDB: a central resource for the superfamily of major intrinsic proteins

- Pune University, Department of Bioinformatics, Pune, India
  Summer student (2007)
  Supervisor: Dr. Indira Ghosh
  Project: Method for mapping active sites of proteins

PEER-REVIEWED PUBLICATIONS

- **Vatsal Agarwal,** Michael Remmert**,** Andreas Biegert**,** Johannes Söding, PDBalert: automatic, recurrent remote homology tracking and protein structure prediction, BMC Structural Biology 2008, 8:51

- Anjali Bansal Gupta, Ravi Kumar Verma, **Vatsal Agarwal**, Manu Vajpai, Vivek Bansal and Ramasubbu Sankararamakrishnan, MIPModDB: a central resource for the superfamily of major intrinsic proteins, Nucleic Acid Research 2012, 40(D1):D362-9

TEACHING EXPERIENCE

- Johns Hopkins University, Baltimore, MD, USA
  Graduate Teaching Assistant
  - Molecules & Cells (Fall 2011)
  - Systems Bioengineering III (Fall 2012)
  - Systems Bioengineering Lab (Spring 2012 & Spring 2013)

PROFESSIONAL EXPERIENCE

- **Tata Consultancy Services Limited,** Noida, India (*2009-2011)*
  Assistant System Engineer
  - Contributed to development and implementation of TCS InstantApps Technology which provides GUI for rapid prototyping of J2EE applications.
  - Created prototype applications for several companies including General Motors**,** CitiBank etc.
  - Coordinated in the development of issue tracking system for Passport Seva project of Indian government
  - Performed regular maintenance of Quantas Airlines ticket system